

Artículo/Article
Sección/Section A

Una alternativa a *Stata*: usando *R* para estimación de Modelos de Regresión

Bolívar Morales-Oñate[1], Carlos Jiménez-Mosquera[2], Paúl Méndez[3]

[1] Facultad de Ciencias, Escuela Superior Politécnica de Chimborazo, Riobamba
[2] Colegio de Ciencias e Ingenierías, Universidad San Francisco de Quito, Quito
[3] Departamento de Ingeniería Matemática, Universidad de Concepción, Concepción
*Autor para correspondencia / Corresponding author bolivar.morales@espoch.edu.ec, cjimenez@usfq.edu.ec, pmendez@ing-mat.udec.cl

An alternative to *Stata* Using *R* for estimation of Regression Models

Resumen

RegUtils es un paquete implementado para el entorno y lenguaje de programación R, que contiene un conjunto de funciones destinadas a facilitar el uso de este software como una alternativa al software comercial *Stata* para la estimación de Modelos de Regresión. Las funciones aquí implementadas, están pensadas para facilitar la migración de modelos estimados en *Stata* al software R, utilizando una sintaxis en muchos sentidos equivalente a la de los comandos de *Stata*, pero que al mismo tiempo sea compatible con paquetes relevantes de R y los paradigmas de programación en R.

Palabras clave: Paquete *RegUtils*, *Stata*, R, migración de modelos.

Código JEL: C13,C87.



Abstract

RegUtils is a package implemented for the environment and programming language R, which contains a set of functions to facilitate the use of this software as an alternative to commercial software Stata for estimating regression models. Functions within this package are designed to help with the migration of models developed in Stata to R. These are constructed keeping a syntax equivalent to Stata counterparts, while maintaining compatibility with R relevant packages and programming paradigms.

Keywords: RegUtils package, Stata, R, migration models.

Editado por / Edited by: Eva O.L. Lantsoght

Recibido / Received: 17/08/2018

Aceptado / Accepted: 06/12/2018

Publicado en línea /PDF online first: 30/01/2019
PDF final: 09/05/2019

INTRODUCCIÓN

El análisis estadístico en general, y en particular, la estimación de regresiones lineales y no lineales, como muchas otras áreas, se apoya en la actualidad en el uso de herramientas computacionales para este análisis. Dos de las herramientas más usadas son: *R* y *Stata*. Estas facilitan, entre otras cosas, el uso de procesos de estimación de parámetros que de otra forma serán inalcanzables, por la complejidad e intensidad del cálculo involucrado. Trabajos como [1], ilustran este punto en el contexto del manejo de grandes bases de datos multinivel longitudinales.

En muchos problemas existe una relación inherente entre dos o más vari- ables, y resulta necesario explorar la naturaleza de esta relación. El análisis de regresión es una t´ecnica estadística para el modelado y la investigación de la relación entre dos o más variables. Unas aplicaciones generales serían: líneas de tendencia en series de tiempo, relaciones causales en medicina, mod- elos econom´etricos, entre otros [2].

R es un lenguaje y entorno de programación para análisis estadístico y gráfico. Se trata de un proyecto de software libre, resultado de la implementación GNU del premiado lenguaje S. R y S-Plus –versión comercial de S– son, probablemente, los dos lenguajes más utilizados en investigación por la comunidad estadística, siendo además muy populares en el campo de la investigación biomédica, la bioinformática y las matemáticas financieras. A esto contribuye la posibilidad de cargar diferentes bibliotecas o paquetes con finalidades específicas de cálculo o procedimientos gráficos [3].

Stata es un paquete de software estadístico creado en 1985 por StataCorp. Es utilizado principalmente por instituciones académicas y empresariales dedicadas a la investigación, especialmente en economía, sociología, ciencias políticas, biomedicina y epidemiología. *Stata* permite, entre otras funcionalidades, la gestión de datos, el análisis estadístico, el trazado de gráficos y las simulaciones [4].

El objetivo principal de este trabajo es describir la implementación y sintaxis de un conjunto de comandos paralelos a la oferta de Stata, a fin de facilitar el uso de R para quienes están familiarizados o vienen de un entorno en el que se usaba principalmente Stata. Estos comandos se escogieron luego de analizar la oferta de modelos de regresión tanto en Stata como en R. La sección 2 de este trabajo ilustra el modelo teórico de cada procedimiento contenido en RegUtils. Luego, en la sección 3 se describe brevemente los pasos para la instalación y uso de RegUtils. La sección 4 presenta los resultados de la implementación de RegUtils donde se evidencia la equivalencia entre R y Stata. Finalmente, el trabajo concluye presentando las conclusiones y posibles desarrollos posteriores.

MARCO TEÓRICO

Se han revisado 35 comandos *Stata* (ver anexo 1), 5 de ellos (14%) no tu- vieron una equivalencia directa en R. Estos son: `areg`, comando que aborda los modelos de regresiones lineales de grandes conjuntos en variables binarias. `boxcox`, es un comando que maneja modelos de regresión con sus transformaciones. `eivreg` se utiliza para modelos de regresión de errores en las variables. `etregress`, maneja modelos de regresión lineal con efectos de tratamiento endógeno. `ivtobit` se usa para la estimación de modelos de regresión *tobit* con variables endógenas.

doi: http://dx.doi.org/10.18272/aci.v11i2.1212



Las siguientes subsecciones están organizadas de tal manera que el lector pueda familiarizarse con la descripción y el modelo asociado a cada comando implementado en `RegUtils`.

Regresión Lineal con conjuntos grandes de variables binarias (`areg`)

Descripción

Este modelo se utiliza cuando se desea ajustar un modelo de regresión lineal que tenga factores dentro del conjunto de variables explicativas de modo que el ajuste implique el uso de un gran número de variables binarias al momento de generar la matriz de diseño.

Modelo

Suponga que desea estimar una regresión con un gran número de variables binarias a partir de *n* individuos:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{d_1}\gamma_1 + \mathbf{d_2}\gamma_2 + \ldots + \mathbf{d_k}\gamma_k + \boldsymbol{\epsilon} \tag{1}$$

donde y es de dimensión n × 1, X (de dimensión $n \times p$ donde p es el número de columnas de la matriz de diseño) es la matriz de variables explicativas (sin incluir las variables binarias), d_i son las variables binarias, $\boldsymbol{\gamma}_i$ son los coeficientes de las variables binarias y $\boldsymbol{\epsilon}$ es el error [5].

Modelos de Regresión con transformaciones de Box Cox (`boxcox`)

Descripción

Este modelo se utiliza para estimar –usando el método de máxima verosimilitud– los parámetros de regresión de variables a las cuales se les ha aplicado la transformación de Box-Cox.

Modelo

En su trabajo seminal, [6] proponen una transformación de los datos que es útil para reducir el sesgo, estabilizar la varianza y concebir que los datos tengan una distribución más parecida a una normal, entre otras [7]. De entre los modelos *lhsonly*, *rhsonly*, *lambda* y *theta* que ofrece Box-Cox, para el ejemplo se utilizó el modelo theta:

$$y_j^{(\theta)} = \beta_0 + \beta_1 x_{1j}^{(\lambda)} + \beta_2 x_{2j}^{(\lambda)} + \cdots + \beta_k x_{kj}^{(\lambda)} + \gamma_1 z_{1j} + \gamma_2 z_{2j} + \cdots + \gamma_l z_{lj} + \epsilon_j \tag{2}$$

donde $\epsilon \sim N(0, \sigma^2)$, y está sujeta a una transformación Box-Cox con parámetro θ y cada $x_1, x_2, \ldots, x_k$ son transformadas por Box-Cox con parámetro λ. Las variables z_1, z_2, . . . z_l son covariables no transformadas [4].

Modelos de regresión con errores en las variables (`eivreg`)

Descripción

Este modelo implica el ajuste de regresiones en las cuales una o más variables independientes son medidas con ruido aditivo. El sesgo introducido por este ruido trata de ser compensando mediante el uso de un coeficiente de confiabilidad.

Modelo

Si una covariable del modelo tiene un error de medición, una regresión tradicional no estimaría adecuadamente su efecto. Además, los coeficientes de las demás covariables del modelo podrían estar sesgados debido a la presencia en el modelo de esa variable. Se puede ajustar el sesgo si se conoce la confiabilidad (*reliability*):

$$reliability = 1 - \frac{noise\,variance}{total\,variance}. \qquad (3)$$

Esto es, dado el modelo

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u} \qquad (4)$$

para alguna variable $\mathbf{x}_i$ en $\mathbf{X}$, $\mathbf{x}_i$ es observada con error, $\mathbf{x}_i = \mathbf{x}_i^* + \mathbf{e}$ y la *varianza de ruido* es la varianza de $\mathbf{e}$. La varianza total es la varianza de $\mathbf{x}_i$ [4]

Modelos de regresión lineal con efectos de tratamiento endógeno (etregress)

Descripción

Este modelo estima los parámetros de una regresión aumentada con una variable binaria endógena, donde el principal objetivo generalmente es parámetro que captura el efecto promedio en el tratamiento (*average treatment effect - ATE*). Esta variable endógena debe estar correlacionada con el tratamiento pero no con el error ni las variables explicativas del modelo principal.

Modelo

El modelo de regresión de efectos de tratamiento endógeno está compuesto de una ecuación para el resultado y_j y una ecuación para el tratamiento endógeno t_j,

$$y_j = \mathbf{x_j}\boldsymbol{\beta} + \delta t_j + \epsilon_j \qquad (5)$$

$$t_j = \begin{cases} 1 & si\ \mathbf{w_j}\boldsymbol{\gamma} + u_j > 0 \\ 0 & en\,otro\,caso \end{cases} \qquad (6)$$

Donde $\mathbf{x}_j$ son las covariables del modelo principal, $\mathbf{w}_j$ son covariables usadas para modelar el tratamiento, y los términos de error $\boldsymbol{\epsilon}_j$ y ujtienen una distribución normal bivariada con media cero y matriz de covarianza:

$$\begin{bmatrix} \sigma^2 & \rho\sigma = \lambda \\ \rho\sigma = \lambda & 1 \end{bmatrix} \qquad (7)$$

Las covariables $\mathbf{x}_j$ y $\mathbf{w}_j$ son exógenas, esto es, no están correlacionadas con el término de error [4].



Modelo de regresión *tobit* con variables endógenas (`ivtobit`)

Descripción

Corresponde a modelos *tobit* donde una o más de las variables regresoras son determinadas endógenamente.

Modelo

El modelo es:

$$y^*_{1i} = \mathbf{y}_{2i}\boldsymbol{\beta} + \mathbf{x}_{1i}\boldsymbol{\lambda} + u_i \quad (8)$$

$$\mathbf{y}_{2i} = \mathbf{x}_{1i}\boldsymbol{\Gamma}_\mathbf{1} + \mathbf{x}_{2i}\boldsymbol{\Gamma}_\mathbf{2} + \mathbf{v_i} \quad (9)$$

donde $i = 1......N$; y_{2i} es un vector $1 \times p$ de variables endógenas; x_{1i} es un vector $1 \times k_1$ de variables exógenas; x_{2i} es un vector $1 \times k_2$ de instrumentos adicionales; y la ecuación y_{2i} está escrita en forma reducida. El modelo supone que $(u_i, \mathbf{v}_i) \sim N(\mathbf{0})$. $\boldsymbol{\beta}$ y $\boldsymbol{\lambda}$ son vectores de parámetros instrumentales, $\boldsymbol{\Gamma}_1$ y $\boldsymbol{\Gamma}_2$ son matrices de parámetros de forma reducida. y y^*_{1i} es observada en forma censurada [7].

INSTALACIÓN

Como se ha mencionado, el anexo 1 contiene la información de todos los comandos investigados y sus paralelos entre Stata y R. La tabla 7 muestra los comados que han sido desarrollados exclusivamente en RegUtils.

Tabla 1. Comandos implementados en RegUtils. Fuente: [3] y [4]. Elaboración propia

Comando en *Stata*	Comando en R	Descripción
areg	alm	Ajusta regresiones con variables dummy
boxcox	boxcox.r	Modelos de regresión Box–Cox
eivreg	eivlm	Regresión con errores en las variables
etregress	etreg	Regresión lineal con efectos endógenos
ivtobit	ivtobit	Regresión tobit con variables endógenas

Actualmente el paquete se encuentra en el repositorio Github: https://github.com/bolimorales/RegUtils. Para utilizarlo se deben realizar los siguientes pasos:

1. Dependencias: `Regutils` depende de los paquetes `MaxLik` [8], `Formula` [9], `car` [10], `sandwich` [11] y `censReg` [12]. Por lo tanto, es necesario instalar estos paquetes previamente.
2. Instalar el paquete `devtools` [13]. Entre otras cosas, este paquete permite instalar paquetes disponibles en el repositorio *GitHub*.
3. Ejecutar los siguientes comandos:
   (a) `library(devtools)`
   (b) `install github("bolimorales/RegUtils")`
4. El comando `library(RegUtils)` le permite usar el paquete

Como es usual en los paquetes de R, se puede acceder a la ayuda de cada una de las funciones de la tabla 7. Por ejemplo, para acceder a la ayuda de `etreg` se ejecuta: `help(etreg)`.

ESTIMACIÓN DE MODELOS

A continuación se presentan los resultados obtenidos de cada una de las funciones del paquete `RegUtils`. Note que se mantiene consistencia con el marco teórico. Es decir, cada uno de los modelos presentados en la sección 2 es ilustrado mediante ejemplos y su respectivo contraste entre R y Stata.

Regresión Lineal con conjuntos grandes de variables binarias (*areg*)

Ejemplo

Se dispone de un conjunto de datos que describen las características de 74 autos. Las variables de interés se muestran en la tabla 2:

Tabla 2. Datos para la regresión con variables binarias. Fuente: [4]. Elaboración propia.

Variable	Descripción	Tipo
mpg	millas por galón	Numérica
Weight	peso en libras	Numérica
gear ratio	proporción entre plato y piñon	Numérica
rep78	medida (de 1 al 5) del registro de reparaciones del auto donde 1 es la peor y 5 es la mejor	Factor

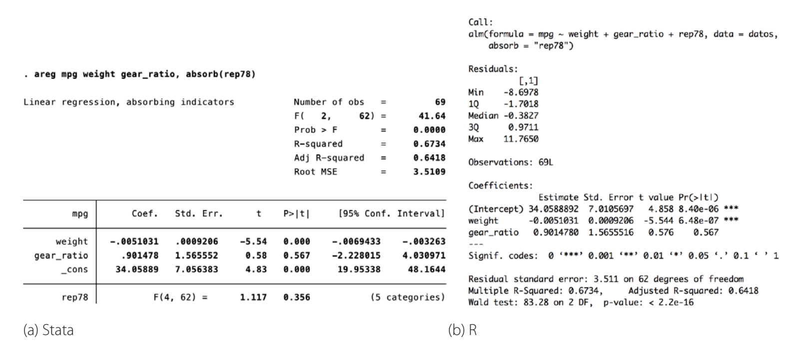
Se desea explicar las millas por galón en función de las demás variables presentadas en la tabla anterior. Una forma de resolver el problema es ajustar un modelo de regresión tradicional, lo cual precisaría de 4 niveles para el factor *rep78*. Sin embargo, `areg` permite, en lugar de realizar una prueba *t* para cada coeficiente, hacer una prueba *F* conjunta para todos los niveles de un determinado factor. Este proceso permite trabajar de forma eficiente sobre conjuntos de datos donde la cantidad de variables binarias generadas es muy grande, pero tambi ́en nos permite centrar el análisis en un conjunto de variables independientes al margen de los grupos creados por las variables binarias [5].

Aplicación

Aquí se ajusta el modelo descrito en la parte anterior tanto en Stata (figura 1a) como en R (figura 1b).



Figura 1. Resultados de la regresión con variables binarias en Stata y R. Fuente y elaboración: autores



```
. areg mpg weight gear_ratio, absorb(rep78)

Linear regression, absorbing indicators          Number of obs   =        69
                                                 F(   2,     62) =     41.64
                                                 Prob > F        =    0.0000
                                                 R-squared       =    0.6734
                                                 Adj R-squared   =    0.6418
                                                 Root MSE        =    3.5109

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0051031   .0009206    -5.54   0.000    -.0069433    -.003263
  gear_ratio |    .901478   1.565552     0.58   0.567    -2.228015    4.030971
       _cons |   34.05889   7.056383     4.83   0.000     19.95338     48.1644
-------------+----------------------------------------------------------------
       rep78 |          F(4, 62) =      1.117   0.356          (5 categories)
```

(a) Stata

```
Call:
alm(formula = mpg ~ weight + gear_ratio + rep78, data = datos,
    absorb = "rep78")

Residuals:
           [,1]
Min    -8.6978
1Q     -1.7018
Median -0.3827
3Q      0.9711
Max    11.7650

Observations: 69L

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.0588892  7.0105697   4.858 8.40e-06 ***
weight      -0.0051031  0.0009206  -5.544 6.48e-07 ***
gear_ratio   0.9014780  1.5655516   0.576    0.567
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.511 on 62 degrees of freedom
Multiple R-Squared: 0.6734,     Adjusted R-squared: 0.6418
Wald test: 83.28 on 2 DF,  p-value: < 2.2e-16
```

(b) R

Note que el comando especifica `absorb(rep78)`, esto indica que se aplica una prueba F sobre el conjunto de los niveles de la variable *rep78*. Su valor p indica que, en conjunto, los niveles de *rep78* no son significativos a un $\alpha = 0.05$. Es decir que los niveles de *rep78* no influyen en las millas por galón. Además, los coeficientes de las demás variables se ajustan en forma estándar.

El resultado puede escribirse de la siguiente manera:

$$\text{mpg} = 34.058 - 0.0051\text{weight} + 0.901\text{gear_ratio}. \qquad (10)$$

Se aprecia que los niveles de la variable *rep78* no se imprimen debido al método utilizado para obtener más eficiencia, es decir, la prueba F. Note que los coeficientes para *weight*, *gear ratio* y la constante son exactamente los mismos al igual que el error estándar, valores t y, consecuentemente, el valor p (ver tablas de coeficientes). En la salida estándar de R no se reportan los intervalos de confianza para los coeficientes pero estos se pueden obtener a través del comando `confint` (usando el modelo guardado en *fit*1) con el que se obtiene los mismos valores (con una pequeña diferencia en el intercepto).

El primer estadístico F (igual a 41.64) mostrado en la salida de Stata corresponde al Chi-cuadrado del test de Wald (83.28) (ubicado al final de la salida de R en la figura 1b), dividido para los grados de libertad, es decir $41.64 = 83.28/2$. En ambos casos el Chi-cuadrado por lo que la conclusión de la prueba es la misma.

También se observa que coinciden el R^2, el R^2 ajustado y los errores estándar residuales (MSE) $= 3.511$, los cuales suelen utilizarse para evaluar la bondad de ajuste de la regresión. La prueba F para *rep78*, no se presenta de forma estándar en R, pero se puede calcular mediante el comando `anova` con el que se obtiene el mismo valor, esto es, 1.117.

Regresión con transformaciones de Box Cox (`box- cox`)

Ejemplo

A continuación se utiliza un subconjunto de datos de la Segunda Encuesta Nacional de Salud y Nutrición (NHANES II), para crear un modelo de estimación del nivel individual de presión arterial. Las variables a utilizar en el modelo se muestran en la tabla 3.

Tabla 3. Datos para la regresión con transformaciones de BoxCox. Fuente: [4]. Elaboración propia

Variable	Descripción	Tipo
bpdiast	Presión arterial	Numérica
bmi	Índice de masa corporal	Numérica
tcresult	Colesterol sérico	Numérica
age	Edad	Numérica
sex	Sexo	Factor

Con el objeto de corregir la no linealidad en la relación con las variables se calcula la transformación de Box-Cox para las variables *bmi*, y *tcresult*, así como la transformación de la variable dependiente dentro del modelo de regresión.

Aplicación

Para estimar este modelo se utiliza en R el comando **boxcox**, parte del paquete RegUtils; en Stata se utiliza el comando **boxcox**. Estos comandos permiten estimar varios tipos de modelos que incluyen el modelo *lambda*, donde se usa el mismo parámetro de transformación sobre todas las variables, modelos *theta*, que utilizan un parámetro diferente para la variable independiente, y modelos que mezclan variables transformadas y no transformadas. Los dos comandos permiten especificar un subconjunto de variables independientes a las cuales no se aplicará la transformación. En la figura 2b se muestran los resultados del ejemplo.



Figura 2. Resultados de la regresión con transformaciones de BoxCox en Stata y R. Fuente y elaboración: autores

```
                                                  Number of obs   =      10351
                                                  LR chi2(5)      =    2515.00
Log likelihood = -39774.987                       Prob > chi2     =      0.000

     bpdiast |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     /lambda |   .6383286   .1577601     4.05   0.000     .3291245    .9475327
      /theta |   .1988197   .0454088     4.38   0.000     .1098201    .2878193

Estimates of scale-variant parameters

             |      Coef.  chi2(df)  P>chi2(df)   df of chi2
-------------+------------------------------------------------
Notrans      |
         age |    .003811   319.060    0.000          1
         sex |  -.1054887   243.284    0.000          1
       _cons |   5.835555
Trans        |
         bmi |   .0872041  1369.235    0.000          1
    tcresult |    .004734    81.177    0.000          1
-------------+------------------------------------------------
      /sigma |   .3348267

   Test            Restricted
    H0:          log likelihood      chi2      Prob > chi2

theta=lambda = -1    -40162.898      775.82       0.000
theta=lambda =  0    -39790.945       31.92       0.000
theta=lambda =  1    -39928.606      307.24       0.000
```

(a) Stata

```
Call:
boxcox.r(formula = bpdiast ~ bmi + tcresult + age + sex, data = datos,
    noTrans = c("age", "sex"))

Box-Cox Estimates :
       Estimate  Std.Err      z  Pr(>|t|)
theta  0.198821 0.047506 4.1852 1.425e-05 ***
lambda 0.638333 0.069578 9.1744 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

LR Test:
X2 = 2515.0, df = 5, P(> X2) = 0.0

Coefficients :
               Estimate      Chi.sq df Chi.sq  Pr(>|t|)
(intercept)  5.8356e+00          NA        NA        NA
bmi          8.7203e-02  1.3692e+03         1 < 2.2e-16 ***
tcresult     4.7339e-03  8.1177e+01         1 < 2.2e-16 ***
age          3.8111e-03  3.1906e+02         1 < 2.2e-16 ***
sex         -1.0549e-01  2.4328e+02         1 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
---
log likehood = -39775.0
```

(b) R

Se escribe los resultados de la siguiente manera:

$$y_j^{(0.198)} = 5.83 + 0.087\text{bmi}^{(0.638)} + 0.0047\text{tcresult}^{(0.638)} + 0.003\text{age} - 0.105\text{sex} \quad (11)$$

Se aprecia que todos los coeficientes son significativos a un $\alpha = 0.05$ como se verifica en la figura 2b. Los coeficientes para lambda y theta coinciden, así como sus pruebas de hipótesis. Se presenta además el mismo número de observaciones, la misma estimación para sigma y de la verosimilitud, −39775. Para realizar la prueba del ratio de verosimilitudes en R se puede usar el comando `logLik.ratio.test` del paquete RegUtils.

Regresión con errores en las variables (`eivreg`)

Ejemplo

Se utiliza datos de una industria automotriz. Se asume que el peso de los autos es medido con ruido aditivo que puede ser aproximado por una confiabilidad de 0.85. Bajo este supuesto se realiza un modelo de regresión lineal simple para el precio utilizando las variables que se detallan en la tabla 4:

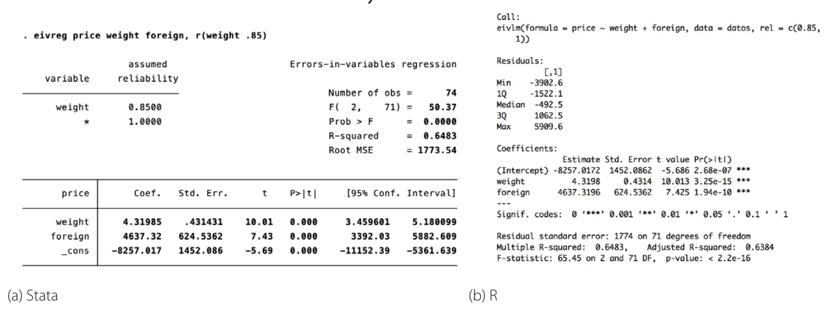
Tabla 4. Datos para la regresión con errores en las variables. Fuente: [4].Elaboración propia.

Variable	Descripción	Tipo
price	Precio	Numérica
weight	peso en libras	Numérica
foreign	Extranjero	Binaria

Aplicación

En Stata el comando `eivreg` ajusta modelos de regresión con errores en las variables. En R se implementó el comando `eivlm` como parte del paquete RegUtils. En la figura 3 se muestran los resultados en Stata y R.

Figura 3. Resultados de la regresión con errores en las variables en Stata y R.
Fuente y elaboración: autores



```
. eivreg price weight foreign, r(weight .85)

                 assumed                    Errors-in-variables regression
    variable    reliability
  ----------------------------                     Number of obs =        74
      weight      0.8500                           F(  2,     71) =    50.37
           *      1.0000                           Prob > F       =   0.0000
                                                   R-squared      =   0.6483
                                                   Root MSE       =  1773.54

  ------------------------------------------------------------------------------
         price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
  -------------+----------------------------------------------------------------
        weight |    4.31985    .431431    10.01   0.000     3.459601    5.180099
       foreign |    4637.32   624.5362     7.43   0.000      3392.03    5882.609
         _cons |  -8257.017   1452.086    -5.69   0.000    -11152.39   -5361.639
  ------------------------------------------------------------------------------
```

(a) Stata

```
Call:
eivlm(formula = price ~ weight + foreign, data = datos, rel = c(0.85,
    1))

Residuals:
          [,1]
Min    -3902.6
1Q     -1522.1
Median  -492.5
3Q      1062.5
Max     5909.6

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -8257.0172  1452.0862  -5.686 2.68e-07 ***
weight          4.3198     0.4314  10.013 3.25e-15 ***
foreign      4637.3196   624.5362   7.425 1.94e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1774 on 71 degrees of freedom
Multiple R-squared:  0.6483,    Adjusted R-squared:  0.6384
F-statistic: 65.45 on 2 and 71 DF,  p-value: < 2.2e-16
```

(b) R

Se puede escribir los resultados de la siguiente manera:

$$\text{price} = -8257.01 + 4.319\text{weight} + 4637.3\text{foreign}. \qquad (12)$$

Los coeficientes habrían sido menores de no haber tomado en cuenta la medición con error de la variable *weight* (Ver [14] para más detalles de este tipo de modelos). El coeficiente de confiabilidad afecta el coeficiente y el error estándar obtenidos para la variable peso. Además se observa que los coeficientes de los parámetros y las pruebas de hipótesis coinciden.

Modelos de regresión lineal con efectos de tratamiento endógeno (`etregress`)

Ejemplo

Se utiliza un subconjunto de la base de datos sobre ingresos de mujeres de Estados Unidos en 1972 con edades entre 18 y 30 años, para modelar los efectos promedios en el tratamiento para la variable union (pertenece o no a un sindicato) sobre el ingreso. Las variables a ser incluidas en el modelo se detallan en la tabla 5.

Tabla 5. Datos para la regresión lineal con efectos de tratamiento endógenos. Fuente: [4].

Variable	Descripción	Tipo
wage	Ingreso	Numérica
grade	Años de instrucción	Numérica
smsa	Variable indicativa de pertenencia a un distrito estadístico	Binaria
black	Variable indicativa para Afro-Americanos	Binaria
tenure	Permanencia en el trabajo actual	Numérica
south	Variable indicativa para residencia en el sur	Binaria

Elaboración propia

Artículo/Article
Sección/Section A
11 (17), 8 – 23

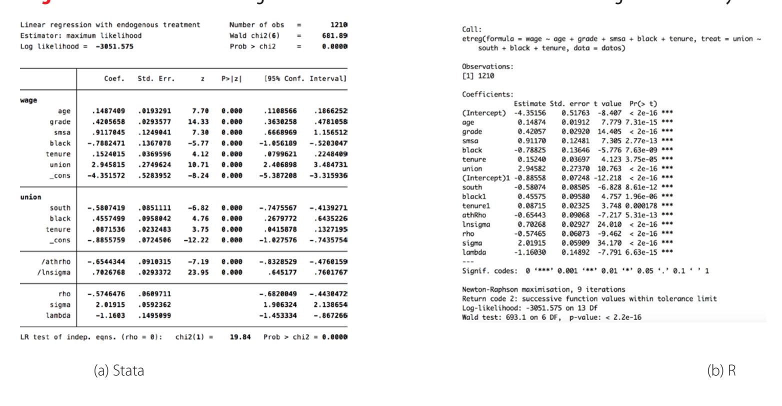


De estas variables se utiliza *south*, *black* y *tenure* para modelar la variable endógena unión.

Aplicación

Para estimar el modelo se emplea el comando `etregress` en Stata y el comando `etreg` del paquete RegUtils en R. En la figura 5 se muestran los resultados obtenidos.

Figure 4. Resultados de la regresión lineal con efectos de tratamiento endógenos en Stata y R.



```
Linear regression with endogenous treatment       Number of obs   =      1210
Estimator: maximum likelihood                     Wald chi2(6)    =    681.89
Log likelihood =  -3051.575                       Prob > chi2     =    0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
wage         |
         age |   .1487409   .0193291     7.70   0.000     .1108566    .1866252
       grade |   .4205658   .0293577    14.33   0.000     .3630258    .4781058
        smsa |   .9117045   .1249041     7.30   0.000     .6668969    1.156512
       black |  -.7882471   .1367078    -5.77   0.000    -1.056189   -.5203047
      tenure |   .1524015   .0369596     4.12   0.000     .0799621    .2248409
       union |   2.945815   .2749624    10.71   0.000     2.406898    3.484731
       _cons |  -4.351572   .5283952    -8.24   0.000    -5.387208   -3.315936
-------------+----------------------------------------------------------------
union        |
       south |  -.5807419   .0851111    -6.82   0.000    -.7475567   -.4139271
       black |   .4557499   .0958042     4.76   0.000     .2679772    .6435226
      tenure |   .0871536   .0232483     3.75   0.000     .0415878    .1327195
       _cons |  -.8855759   .0724506   -12.22   0.000    -1.027576   -.7435754
-------------+----------------------------------------------------------------
     /athrho |  -.6544344   .0910315    -7.19   0.000    -.8328529   -.4760159
    /lnsigma |   .7026768   .0293372    23.95   0.000      .645177    .7601767
-------------+----------------------------------------------------------------
         rho |  -.5746476   .0609711                     -.6820049   -.4430472
       sigma |    2.01915   .0592362                      1.906324    2.138654
      lambda |    -1.1603   .1495099                     -1.453334    -.867266
------------------------------------------------------------------------------
LR test of indep. eqns. (rho = 0):   chi2(1) =    19.84   Prob > chi2 = 0.0000
```

(a) Stata

```
Call:
etreg(formula = wage ~ age + grade + smsa + black + tenure, treat = union ~
    south + black + tenure, data = datos)

Observations:
[1] 1210

Coefficients:
              Estimate Std. error t value  Pr(> t)
(Intercept)   -4.35156    0.51763  -8.407  < 2e-16 ***
age            0.14874    0.01912   7.779 7.31e-15 ***
grade          0.42057    0.02920  14.405  < 2e-16 ***
smsa           0.91170    0.12481   7.305 2.77e-13 ***
black         -0.78825    0.13646  -5.776 7.63e-09 ***
tenure         0.15240    0.03697   4.123 3.75e-05 ***
union          2.94582    0.27370  10.763  < 2e-16 ***
(Intercept)1  -0.88558    0.07248 -12.218  < 2e-16 ***
south         -0.58074    0.08505  -6.828 8.61e-12 ***
black1         0.45575    0.09580   4.757 1.96e-06 ***
tenure1        0.08715    0.02325   3.748 0.000178 ***
athRho        -0.65443    0.09068  -7.217 5.31e-13 ***
lnsigma        0.70268    0.02927  24.010  < 2e-16 ***
rho           -0.57465    0.06073  -9.462  < 2e-16 ***
sigma          2.01915    0.05909  34.170  < 2e-16 ***
lambda        -1.16030    0.14892  -7.791 6.63e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Newton-Raphson maximisation, 9 iterations
Return code 2: successive function values within tolerance limit
Log-likelihood: -3051.575 on 13 Df
Wald test: 693.1 on 6 DF,  p-value: < 2.2e-16
```

(b) R

Fuente y elaboración: autores

El modelo ajustado del ejemplo sería:

$$\text{wage} = -4.35+0.148\text{age}+0.420\text{grade}+0.911\text{smsa}-0.788\text{black}+0.152\text{tenure}+2.94\text{union} \quad (13)$$

$$\text{union} = \begin{cases} 1 & si \; -0.885 - 0.580\text{south} + 0.455\text{black} + 0.087\text{tenure} > 0 \\ 0 & \text{en otro caso} \end{cases} \quad (14)$$

$$\begin{bmatrix} 2.01^2 & -0.574 * 2.01 = -1.160 \\ -0.574 * 2.01 = -1.160 & 1 \end{bmatrix} \quad (15)$$

Se observa en la figura 8 que los coeficientes son significativos y que el efecto de tratamiento endógeno (coeficiente de union) es 2.94

Modelo de regresión *tobit* con variables endógenas (`ivtobit`)

Ejemplo

Se utiliza a continuación datos de ingresos de un grupo de mujeres asumiendo que todas las mujeres que deciden no trabajar reciben $ 10,000 en pagos de asistencia social y la manutención de los hijos. En el modelo se incluirán las variables que se muestran en la tabla 6:

Tabla 6. Datos para la regresión tobit con variables endógenas. Fuente: [4]. Elaboración propia.

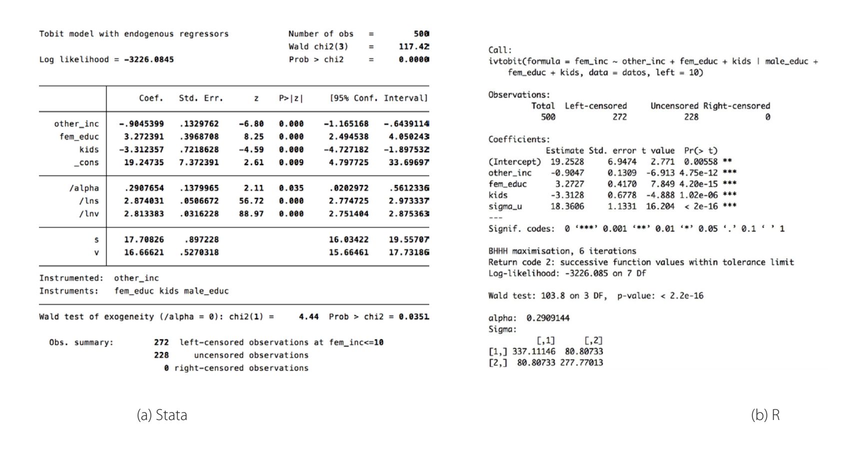
Variable	Descripción	Tipo
fem_inc	Ingreso	Numérica
fem_educ	Años de instrucción formal	Numérica
kids	Número de hijos	Numérica
other_inc	Otros ingresos del hogar	Numérica
male_educ	Años de instrucción formal del esposo	Numérica

En el modelo se considera la varibale *other_inc* como endógena, motivo por la cual se utiliza la variable *male_educ* como variable instrumental.

Aplicación

Para estimar el modelo se utiliza el comando ivtobit de Stata. En R se usa el comando del mismo nombre implementado como parte del paquete RegUtils. La figura 5 muestra los resultados obtenidos.

Figura 5. Resultados de la regresión tobit con variables endógenas en Stata y R. Fuente y elaboración: autores



```
Tobit model with endogenous regressors          Number of obs   =        500
                                                Wald chi2(3)    =     117.42
Log likelihood = -3226.0845                     Prob > chi2     =     0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   other_inc |  -.9045399    .1329762    -6.80   0.000    -1.165168   -.6439114
    fem_educ |   3.272391    .3968708     8.25   0.000     2.494538    4.050243
        kids |  -3.312357    .7218628    -4.59   0.000    -4.727182   -1.897532
       _cons |   19.24735    7.372391     2.61   0.009     4.797725    33.69697
-------------+----------------------------------------------------------------
      /alpha |   .2907654    .1379965     2.11   0.035     .0202972    .5612336
        /lns |   2.874031    .0506672    56.72   0.000     2.774725    2.973337
        /lnv |   2.813383    .0316228    88.97   0.000     2.751404    2.875363
-------------+----------------------------------------------------------------
           s |   17.70826     .897228                      16.03422    19.55707
           v |   16.66621    .5270318                      15.66461    17.73186
------------------------------------------------------------------------------
Instrumented:  other_inc
Instruments:   fem_educ kids male_educ
------------------------------------------------------------------------------
Wald test of exogeneity (/alpha = 0): chi2(1) =     4.44  Prob > chi2 = 0.0351

  Obs. summary:        272  left-censored observations at fem_inc<=10
                       228     uncensored observations
                         0 right-censored observations
```

(a) Stata

```
Call:
ivtobit(formula = fem_inc ~ other_inc + fem_educ + kids | male_educ +
    fem_educ + kids, data = datos, left = 10)

Observations:
         Total  Left-censored     Uncensored Right-censored
           500            272            228              0

Coefficients:
            Estimate Std. error t value  Pr(> t)
(Intercept)  19.2528     6.9474   2.771  0.00558 **
other_inc    -0.9047     0.1309  -6.913 4.75e-12 ***
fem_educ      3.2727     0.4170   7.849 4.20e-15 ***
kids         -3.3128     0.6778  -4.888 1.02e-06 ***
sigma_u      18.3606     1.1331  16.204  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

BHHH maximisation, 6 iterations
Return code 2: successive function values within tolerance limit
Log-likelihood: -3226.085 on 7 Df

Wald test: 103.8 on 3 DF,  p-value: < 2.2e-16

alpha:  0.2909144
Sigma:
          [,1]      [,2]
[1,] 337.11146  80.80733
[2,]  80.80733 277.77013
```

(b) R

El modelo estimado es:

$$\text{fem_inc} = 19.24 - 0.9045\text{other_inc} + 3.27\text{fem_educ} - 3.31\text{kids} \qquad (16)$$

En ambos comandos el estimador se calcula utilizando el método de máxima verosimilitud. Todas las variables del modelo son estadísticamente significativas a un $\alpha = 0.05$ y se aprecia que son los mismos valores.



CONCLUSIONES

El uso del paquete RegUtils, sin duda es una herramienta muy útil para usuarios que desean migrar o usar simultáneamente R y Stata dado que permite obtener modelos similares a aquellos ajustados con los comandos de Stata, con una sintaxis que resulta muy fácil de equiparar con la de Stata.

El paquete facilita que estos modelos se integren con otras funcionalidades y comandos de R. Sin embargo, se recomienda que una vez que se alcance el nivel adecuado de familiaridad con R, se explore otros paquetes y rutinas propias del lenguaje que, sin tener las mismas salidas de Stata pueden conducirnos al ajuste de modelos de utilidad.

Utilizar los comandos implementados en R para el manejo de estos modelos conlleva algunas ventajas sobre el software propietario, se podría mencionar entre estas el hecho de que no tendría costo para el usuario. Además, R es versátil para el desarrollo de rutinas, acceso al código para aprender de él y modificarlo en función de las necesidades del investigador y facilidades para el trabajo colaborativo al poder compartir análisis sin preocupación de que el receptor disponga de las licencias.

Como todos los proyectos realizados en R, se recomienda que todo usuario interesado se motive en generar retroalimentación para su perfeccionamiento. Por ejemplo podría colaborar con paquetes o correcciones de paquetes existentes en el Comprehensive R Archive Network – CRAN. También se puede contribuir a mejorar las rutinas del paquete RegUtils, ya que el código se encuentra disponible en el sitio github.

REFERENCES

[1] D. F. McCaffrey, J. R. Lockwood, K. Mihaly, T. R. Sass, et al., *A review of stata commands for fixed-effects estimation in nor- mal linear models*, Stata Journal, 12 (2012), p.406.

[2] B. Baltagi, *Econometric analysis of panel data*, John Wiley & Sons, 2002.

[3] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017.

[4] StataCorp, *Stata 15 Base Reference Manual*, Stata Press, College Station, TX, 2017.

[5] M. C. Lovell, *A simple proof of the fwl (frisch-waugh-lovell) theorem*. 2006.

[6] G. E. Box and D. R. Cox, *An analysis of transformations*, Journal of the Royal Statistical Society. Series B (Methodological), (1964), pp. 211– 252.

[7] J. M. *Wooldridge, Introducción a la econometr´ıa: un enfoque mod- erno*, Editorial Paraninfo, 2006.

[8] A. Henningsen and O. Toomet, *maxlik: A package for maximum likelihood estimation in R*, Computational Statistics, 26 (2011), pp.443– 458.

[9] A. Zeileis and Y. Croissant, *Extended model formulas in R: Mul- tiple parts and multiple responses*, Journal of Statistical Software, 34 (2010), pp. 1–13.

[10] J. Fox and S. Weisberg, *An R Companion to Applied Regression*, Sage, Thousand Oaks CA, second ed., 2011.

[11] A. Zeileis, *Econometric computing with hc and hac covariance matrix estimators*, Journal of Statistical Software, 11 (2004).

[12] A. Henningsen, *censReg: Censored Regression (Tobit) Models*, 2017. R package version 0.5-26.

[13] H. Wickham and W. Chang, *devtools: Tools to Make Developing R Packages Easier*, 2017. R package version 1.13.4.

[14] N. R. Draper and H. Smith, *Applied regression analysis*, vol. 326, John Wiley & Sons, 2014.

[15] B. Morales-Oñate, C. Jiménez-Mosquera, P. Mendez, and V. Morales-Oñate, *RegUtils: Tools for STATA users in Regression Models Estimation*, 2015. R package version 0.1.

[16] A. Ghalanos, *rugarch: Univariate GARCH models.*, 2018. R package version 1.4-0.

[17] D. Wuertz, T. Setz, and Y. Chalabi, *fArma: Rmetrics - Modelling ARMA Time Series Processes*, 2017. R package version 3042.81.

[18] T. Coelli and A. Henningsen, *frontier: Stochastic Frontier Analy- sis*, 2017. R package version 1.1-2.

[19] P. Chaussé, *Computing generalized method of moments and generalized empirical likelihood with R*, Journal of Statistical Software, 34 (2010), pp. 1–35.

[20] O. Toomet and A. Henningsen, *Sample selection models in R: Package sampleSelection*, Journal of Statistical Software, 27 (2008).

[21] O. Toomet, *intReg: Interval Regression*, 2015. R package version 0.2-8.

[22] C. Kleiber and A. Zeileis, *Applied Econometrics with R*, Springer- Verlag, New York, 2008. ISBN 978-0-387-77316-2.

[23] A. Henningsen and J. D. Hamann, *systemfit: A package for es- timating systems of simultaneous equations in r*, Journal of Statistical Software, 23 (2007), pp. 1–40.

[24] R. Koenker, *quantreg: Quantile Regression*, 2017. R package version 5.34.





[25] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, Springer, New York, fourth ed., 2002. ISBN 0-387-95457-0.

[26] Y. Rosseel, *lavaan: An R package for structural equation modeling*, Journal of Statistical Software, 48 (2012), pp. 1–36.

[27] Y. Croissant and A. Zeileis, *truncreg: Truncated Gaussian Regres- sion Models*, 2016. R package version 0.2-4.

[28] G. Millo, *Robust standard error estimators for panel models: A uni- fying approach*, Journal of Statistical Software, 82 (2017), pp. 1–27.

[29] K. Kashin, *panelAR: Estimation of Linear AR(1) Panel Data Mod- els with Cross-Sectional Heteroskedasticity and/or Correlation*, 2014. R package version 0.1.

[30] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, and R Core Team, *nlme: Linear and Nonlinear Mixed Effects Models*, 2017. R package version 3.1-131.

ANEXO

Tabla 7. Cuadro comparativo de comandos entre Stata y R. Fuente: [20] y [22]. Elaboración propia

Comando en Stata	Librería	Comando en R	Descripción
areg	RegUtils[18]	alm	Ajusta regresiones con variables dummy
arch	rugarch[8]	ugarchFit	Modelos de regresión con errores ARCH
arima	fArma[28]	armaFit	Modelos ARIMA
boxcox	RegUtils	boxcox.r	Modelos de regresión Box?Cox
cnsreg	stats[20]	lm	Regresión lineal restringida
eivreg	RegUtils	eivlm	Regresión con errores en las variables
etregress	RegUtils	etreg	Regresión lineal con efectos endógenos
frontier	frontier[4]	sfa	Modelos estocásticos frontier
gmm	gmm[3]	gmm	método generalizado de momentos de estimación
heckman	sampleSelection[24]	selection	Modelos de selección Heckman
intreg	intReg[23]	intReg	Regresión intervalo
ivregress	AER[13]	ivreg	Regresión de variables instrumentales simples
ivtobit	RegUtils	ivtobit	Regresión tobit con variables endógenas
newey	sandwich[29]	NeweyWest	Regresión de Newey ?West con errores estándar
nl	stats	nls	Estimación no lineal de mínimos cuadrados
nlsur	systemfit[10]	nlsystemfit	Estimación de sistemas de ecuaciones no lineales
qreg	quantreg[14]	rq	Regresión cuantil (incluye mediana)
reg3	systemfit	systemfit	Regresión de mínimos cuadrados en tres estados (3SLS)
rreg	MASS[25]	rlm	Un tipo de regresión robusta
gsem	lavaan[21]	cfa	modelos de ecuaciones estructurales generalizadas
sem	lavaan	cfa	modelos de ecuaciones estructurales lineales
sureg	systemfit	systemfit	regresión aparentemente no relacionada
tobit	censReg[9]	censReg	Regresión tobit
truncreg	truncreg[5]	truncreg	Regresión truncada
xtabond	plm[17]	pgmm	Estimación de datos de panel dinámicos de Arellano?Bond
xtdpd	plm	pgmm	Estimación de datos de panel dinámicos lineales
xtfrontier	frontier	sfa	Modelos frontier de datos de panel dinámicos
xtgls	panelAR[12]	panelAR	Modelos GLS de datos de panel
xthtaylor	plm	pht	Modelos de estimación de errores de Hausman?Taylor
xtintreg	plm	survreg	Modelos de regresión de datos de panel en intervalos
xtivreg	plm	plm	Regresión de datos de panel con variables instrumentales (2SLS)
xtpcse	panelAR	panelAR	Regresión lineal con errores estándar
xtreg	plm	plm	Modelos lineales con efectos aleatorios
xtregar	nlme[19]	lme	Modelos lineales con efectos aleatorios AR(1)
xttobit	censReg	censReg	Modelos tobit de datos de panel